Tools for Grading Evidence: strengths, weaknesses, and the impact on effective knowledge translation

Relevant conference theme: Concepts and methods

Translating Research Into Practice – Dementia and Public Health (TRIP-DPH)

Michelle Irving, Nic Cherbuin, Ranmalee Eramudugolla, Peter Butterworth, Lily O'Donoughue Jenkins, Kaarin Anstey
Rating evidence and knowledge translation: Could grading tools be selling us short?

• Knowledge Translation is defined by the World Health Organization as “the synthesis, exchange, and application of knowledge by relevant stakeholders to accelerate the benefits of global and local innovation in strengthening health systems and improving people’s health” (WHO, 2005).

• This objective is compromised when a body of research is oversimplified, graded incorrectly, or not fully understood by those relying on flawed grading systems to inform their decision making.
TRIP-DPH is an ARC Linkage collaboration between the Australian National University, ACT Health and Alzheimer’s Australia. The project is examining multiple barriers to effective knowledge translation and what can be done to make improvements.

One of the barriers identified pertains to the vast quantities of research that decision makers are confronted with and the issues that arise from trying to evaluate the quality of such evidence. How do those working outside the research world distinguish strong evidence from weak evidence?

This well known problem has broadly been addressed through the development of hundreds of grading instruments for evaluating individual studies and systematic reviews.

If these tools are flawed, over-reliance on them may lead to miscommunication and incorrect assessments of the evidence base, compromising knowledge translation.
What is grading?

- Grading instruments provide a metric to ‘quantify’ the quality of evidence from either an individual study or from a body of evidence.

- Typically, they comprise a list of criteria and instructions for application of these criteria in order to achieve a quality rating or ‘grade’.

- Ratings for scales and checklists are frequently presented as a whole score or fraction (e.g. 18/20) while ratings that are applied to bodies of evidence are usually grade A, B, C etc.
## Examples of grading instruments

<table>
<thead>
<tr>
<th>Grading a body of evidence</th>
<th>Grading individual studies</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grading of Recommendations Assessment, Development and Evaluation (GRADE)</td>
<td>Critical Appraisal Skills Program (CASP)</td>
</tr>
<tr>
<td>National Health and Medical Research Council - FORM (NHMRC-FORM)</td>
<td>Graphic Appraisal Tool for Epidemiology (GATE)</td>
</tr>
<tr>
<td>Scottish Intercollegiate Guidelines Network (SIGN)</td>
<td>Downs and Black Checklist</td>
</tr>
<tr>
<td>National Service Framework for Long Term Conditions (NSF-LTC)</td>
<td>Newcastle-Ottawa Scale (NOS)</td>
</tr>
<tr>
<td>Strength of Recommendations Taxonomy (SORT)</td>
<td>Risk of Bias Assessment Tool for Nonrandomized Studies (RoBANS)</td>
</tr>
<tr>
<td>National Institute for Clinical Evidence (NICE) Guidelines</td>
<td>Jadad Scale (aka Oxford Quality Scoring System)</td>
</tr>
</tbody>
</table>
What types of criteria are included?

While these tools vary, typically the criteria used to rate evidence includes some or all of the below:

Evidence can be rated up or down depending on how well it meets criteria in any one of these domains. Different grading systems place less or more emphasis on different components.

### Why is grading evidence a good idea?

<table>
<thead>
<tr>
<th><strong>Systematic approach</strong></th>
<th><strong>Time effective</strong></th>
<th><strong>Objective</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Can minimise bias in assessments</td>
<td>Instruments often facilitate quick assessments</td>
<td>Not reliant on author(s)’s view of the strength of their research</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Accessible</strong></th>
<th><strong>Clear</strong></th>
<th><strong>High Impact</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Once studies have been graded the results are usually broadly available for others to see and use</td>
<td>Grades are a straightforward way to communicate to those who are not necessarily equipped to evaluate research</td>
<td>Grades of systematic reviews are often used to address real world problems (e.g. clinical guidelines)</td>
</tr>
</tbody>
</table>
What are the problems?

- Lack of validation and poor reliability
- Highly subjective - low interrater reliability
- Create false sense of trust or over-reliance on grading outcomes
- Overly complicated, unclear instructions and little or no training – mistakes made
- Biased in favour of certain types of studies (e.g. randomised control trials) and against others (e.g. observational)
- Grading of evidence may not be applicable to local settings

(Bagshaw & Bellomo, 2008; Baker et al., 2011; Baker et al., 2010; Barbui et al., 2010; Brouwers et al., 2005; Calderón et al., 2006; Colle et al., 2002; Gugiu & Gugiu, 2010; Hartline et al. 2012; Ibargoyen-Roteta et al., 2010; Juni et al., 1999; Katrak et al., 2004; Kavanagh, 2009; Persaud & Mamdani, 2006)
Why should policy makers, practitioners and other decision makers care?

- Even if you don’t directly use one of these instruments, chances are that you have relied on research that is evaluated by them: e.g. NHMRC guidelines, Cochrane Collaboration reviews, WHO recommendations, NIH reports

- This research may have been graded in ways that do not accurately represent the strength of the evidence, leading to flawed conclusions

- Decisions made on the basis of erroneous evidence could lead to resources being misdirected to less worthy causes or away from more worthy causes
Example: Smoking as a risk factor for Alzheimer’s Disease

• In 2010 the US Department of Health and Human Services National Institutes of Health (NIH) released a report, “Preventing Alzheimer’s Disease and Cognitive Decline” (Williams et al., 2010). It aimed to, ‘help clinicians, employers, policymakers, and others make informed decisions about the provision of health care services.’

• In this report a systematic review by Anstey et al. (2007) on the effects of smoking on Alzheimer’s Disease was given a ‘low’ quality rating, as assessed using the GRADE approach. There are at least two major problems with this rating:
Example: Smoking as a risk factor for Alzheimer’s Disease

1. The systematic review was downgraded because it relied on observational studies rather than using randomised control trials. However in research on smoking it is not ethical to perform randomised control trials, as such observational studies were the strongest possible form of evidence available.

2. Grading failed to take into account additional sources of evidence that fall outside the scope of GRADE’s rating criteria. This includes a convergence of evidence from other research areas, such as animal studies that have also linked smoking and cognitive decline.
Potential outcome of flawed grading

If the strongest possible evidence is graded as ‘low’ due to ineffective use of grading systems it is possible that health practitioners and policy makers may conclude that there are no grounds to recommend reduction or cessation of smoking as a preventative means of decreasing dementia. Smokers themselves may also directly interpret such information as reassurance that their use of tobacco is not linked with increased risk. Such an outcome has the potential to maintain the prevalence of this costly disease at higher levels.
Where to from here?

1. At a minimum researchers, policy makers and practitioners should be aware of the shortcomings of grading instruments. Critical evaluation of their conclusions is necessary before relying on grades as a form of evidence for decision making.

2. There should be greater emphasis on applying the correct grading system to the type of evidence.

3. Options include creating a new instrument, having multiple instruments that are more targeted for specific types of research, or modifying existing instruments to address criticisms.


Barbui, C., T. Dua, et al. (2010). Challenges in Developing Evidence-Based Recommendations Using the GRADE Approach: The Case of Mental, Neurological, and Substance Use Disorders. *PLOS Medicine, 7*(8)


Kavanagh, B. P. (2009). The GRADE system for rating clinical guidelines. *PLoS Medicine, 6*(9), e1000094


Grading Instrument References

CASP: Details available at the Critical Appraisal Skills Program website: http://www.casp-uk.net/


Contact author

Michelle Irving
Centre for Research on Ageing, Health and Wellbeing
Australian National University
michelle.irving@anu.edu.au
www.crahw.anu.edu.au

The First Global Conference on Research Integration and Implementation,
http://www.I2Sconference.org/